УДК 004.89:004.4

С.М. Вороной, А.А. Егошина
Государственный университет информатики и искусственного интеллекта,
г. Донецк, Украина
smv@iai.donetsk.ua

Определение грамматических характеристик словоформы методом графов

Для систем автоматизированного перевода и морфологического процессора полнотекстовых информационно-поисковых систем с естественноязыковым интерфейсом предложен алгоритм получения всех характеристик словоформ субстантивного и адъективного склонения с помощью графов составляющих их морфем.

Введение

На современном этапе развития информационных технологий неотъемлемой частью интеллектуальных информационно-поисковых систем (ИПС) и систем машинного перевода является морфологический компонент, представляющий собой комплекс программ, обеспечивающих морфологический анализ и синтез лексем ЕЯ.

В любой морфологической модели, учитывающей значения грамматических характеристик лексем, с каждой лексемой связаны: синтаксический класс (часть речи), словоизменительный (парадигматический) класс и значения грамматических категорий, или грамматических переменных (ГП), соответствующих синтаксическому классу [1].

Если морфологический анализатор работает со словарем словоформ, то задача морфологического анализа сводится к задаче поиска заданной словоформы в базе данных, где с каждой словоформой связаны ее грамматические характеристики. Результатом морфологического анализа словоформы являются грамматические характеристики, связанные с флексией, и начальная форма соответствующей лексемы.

Перечень всех морфологических характеристик слова и их возможные значения зависят от конкретного языка. Хотя некоторые грамматические характеристики, например часть речи, присутствуют во многих языках.

В настоящее время наиболее распространены три подхода к проведению морфологического анализа. Первый подход нахождения грамматических характеристик словоформы опирается на морфологическую модель, представленную в «Грамматическом словаре русского языка» А.А. Зализняка [2].

Второй подход основывается на некоторой системе правил, по заданному слову определяющих его морфологические характеристики. Третий, вероятностный, подход основан на сочетаемости слов с конкретными морфологическими характеристиками. Данный подход в основном применяется для обработки текстов с фиксированным порядком слов в предложении [3].

Постановка задачи

Слова, поступающие на вход модуля морфологического анализа, могут не входить в словарь всех словоформ. Данная ситуация возникает вследствие ошибок, полученных на этапе ввода исходного текста. В таком случае применение метода

морфологического анализа, основанного на использовании словаря А.А. Зализняка, на дает требуемых результатов.

При вероятностном способе [4] проведения морфологического анализа для каждой словоформы определяются все ее грамматические классы, а также вероятность ее отношения к каждому из этих классов. Это выполняется на основе некоторого набора документов, где каждому слову предварительно поставлен в соответствие грамматический класс. После этого вычисляются вероятности сочетаний определенных грамматических классов для слов, стоящих рядом. На основе этих чисел может проводиться анализ слов, но для него необходимо уже не только само слово, но и стоящие рядом с ним слова. Как было замечено выше, вероятностный метод применим только для тех языков, у которых четко фиксирован порядок слов в предложении. Если же порядок слов можно изменять, то все возможные сочетания грамматических классов будут практически равновероятны.

В той ситуации, когда не удалось определить характеристики слова с помощью методов четкой морфологии или порядок слов в исходном предложении не является строго фиксированным, задача морфологического анализа может быть решена с помощью «нечеткой» морфологии. Наличие тех или иных лексем может определять морфологические характеристики слова: можно построить систему правил, которая будет опираться на наличие или отсутствие каких-либо частей и выдавать одно или несколько предположений о морфологических параметрах.

В [5] предложен алгоритм морфологического анализа, основанный на работе со словарями морфем, а не словоформ. **Целью** настоящего исследования является разработка алгоритма получения всех характеристик словоформы с помощью графов составляющих ее морфем.

Определение грамматических характеристик словоформ субстантивного и адъективного склонения

Как русский, так и украинский языки принадлежат к языкам флективно-аналитического типа. Их особенность заключается в том, что грамматические элементы значения передаются, как правило, особыми единицами плана выражения, входящими в состав слова – флексиями.

По флексии можно определить, к той или иной грамматической категории относится лексема. В основу классификации лексем можно положить выраженность одних и тех же морфологических категорий [6].

В работе рассматриваются два типа словоформ: субстантивное и адъективное склонение – и возможности определения и перевода словоформы с одного языка на другой, учитывая значения флексий и словообразовательных морфов.

Строгая закономерность сочетания аффиксов с определенным типом основ позволяет нам получить грамматические характеристики словоформы по составляющим ее морфам.

Так, например, в результате разбиения слова книг/ами на составляющие его аффиксы получаем следующие грамматические характеристики: имя существительное, множественное число, творительный падеж.

Предлагается использовать для получения грамматических характеристик словоформы взвешенный граф переходов. В подобном графе конечные вершины (состояния) будут соответствовать грамматическим характеристикам словоформы, а дуги – значениям аффиксов.

В данной работе рассматривается построение такого графа для словоформ субстантивного и адъективного склонения.

Граф анализа словоформ субстантивного склонения приведен на рис. 1. Начальная вершина р1 соответствует появлению на входе некоторого окончания, значение которого совпадает с одной из дуг, исходящих из вершины р1. Дуги графа имеют вес, равный возможному окончанию словоформы субстантивного склонения. Конечные вершины (11, 12…71, 72) обозначают, что анализируемая словоформа обладает такими грамматическими характеристиками. Первая цифра в обозначении конечной вершины обозначает предшествующее состояние (вершину), а вторая цифра – порядковый номер.

Однако в русском языке существует определенное количество морфем, которые обладают омонимией. Например, рассмотрим слова, в состав которых входит суффикс -ец, который имеет несколько значений.

1. Названия лиц мужского рода:
- обладающие определенными качествами (глупец);
- национальность (китаец).

2. Название предметов (резец).

3. Обозначение имени существительного мужского рода уменьшительно-ласкательного значения (братец, хлебец).

Суффиксальный аффикс, присоединяясь к основам, принадлежащим к определенному семантическому кругу, приносит в слово нужное значение.

Широко распространена омонимия в окончаниях.

На рис. 1 вершинам, обладающим морфологической омонимией, соответствуют вершины р2, р3, р4, р5, р6.

Если омонимия имеет место в окончаниях, то ее можно не учитывать при словоизменении во время перевода с одного флексивного языка на другой.

Если же слово производное, то следует проверить омонимию суффиксов. Суффикс может присоединяться не только к корню, но и к другому суффиксу, который, в свою очередь, присоединен к корневой морфеме. Если в производном слове несколько суффиксов, то только конечный суффикс указывает на ту часть речи, к которой относится словоформа. Так слово учи/тель относится к разряду имен существительных, так как суффикс -тель является суффиксом имени существительного, а слово учи/тель/ск/ий – к разряду имен прилагательных, так как это слово оканчивается суффиксом прилагательного -ск и т.п.

Определяющая роль конечного суффикса особенно ярко выражена в словах типа учи/тель/ниц/а, жи/тель/ниц/а. Хотя суффиксы мужского рода (-тель) в этих словах продолжают оставаться, принадлежность слова к женскому роду определяется конечным суффиксом -ниц/а [6].

Учитывая сказанное, будем считать, что вершины графа, обладающие омонимией, представляют собой некоторые дополнительные подграфы анализа суффиксальных аффиксов словоформы.

Аналогичным образом можно построить граф анализа словоформ адъективного склонения, представленный на рис. 2.

Алгоритм морфологического анализа для производных слов заключается в следующем.

1. Проводится разбиение словоформы на составляющие ее морфемы по алгоритму [5].

2. По суффиксу определяется часть речи, к которой принадлежит анализируемая словоформа.

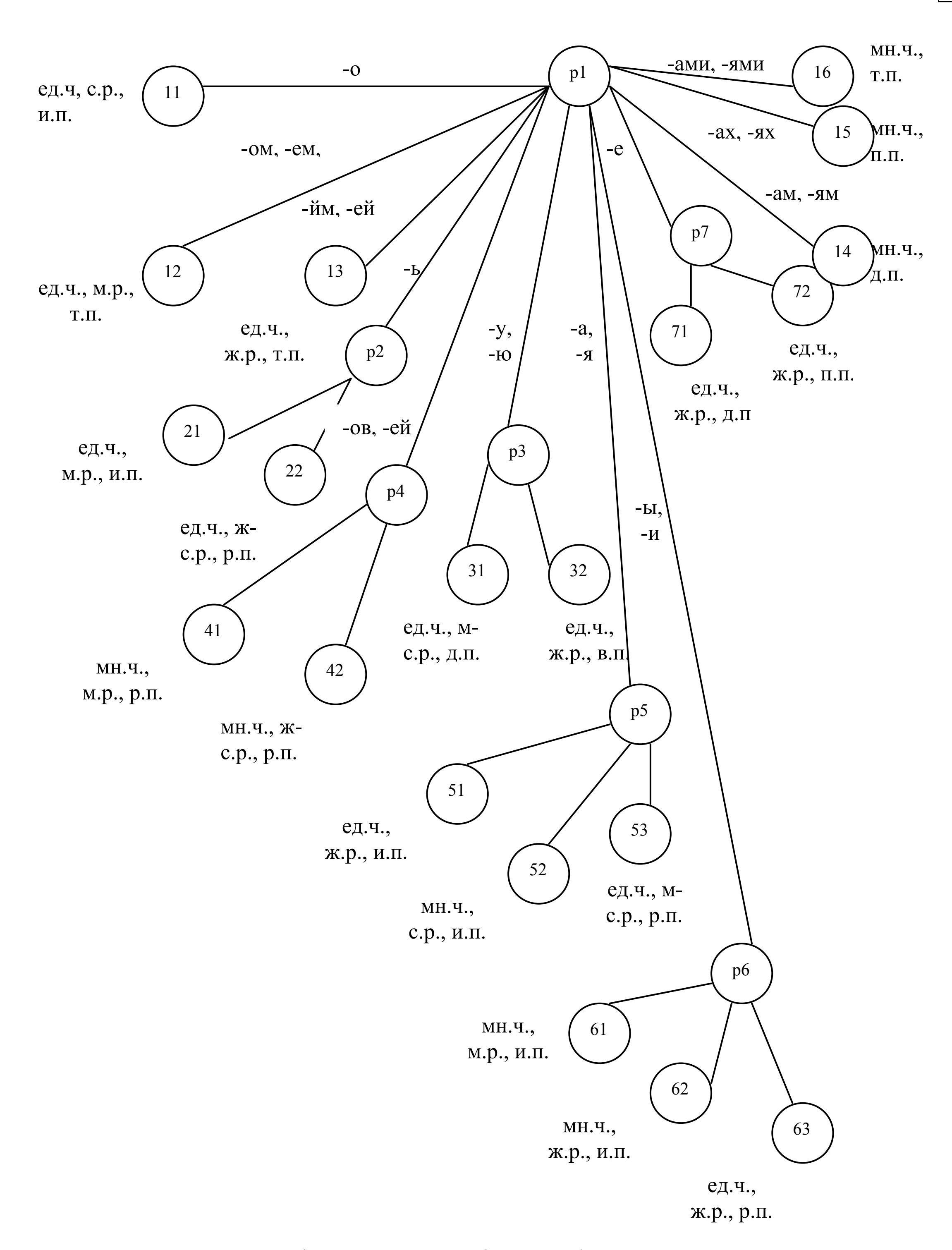


Рисунок 1 – Граф анализа словоформ субстантивного склонения

3. Производится поиск окончания в соответствующем определенной части речи графе.

4. Если найденное окончание не имеет омонимии, то определяются грамматические категории рода, числа и падежа.

5. Если же возникает омонимия, то производится дальнейший анализ с помощью построенного аналогичным образом подграфа суффиксов.

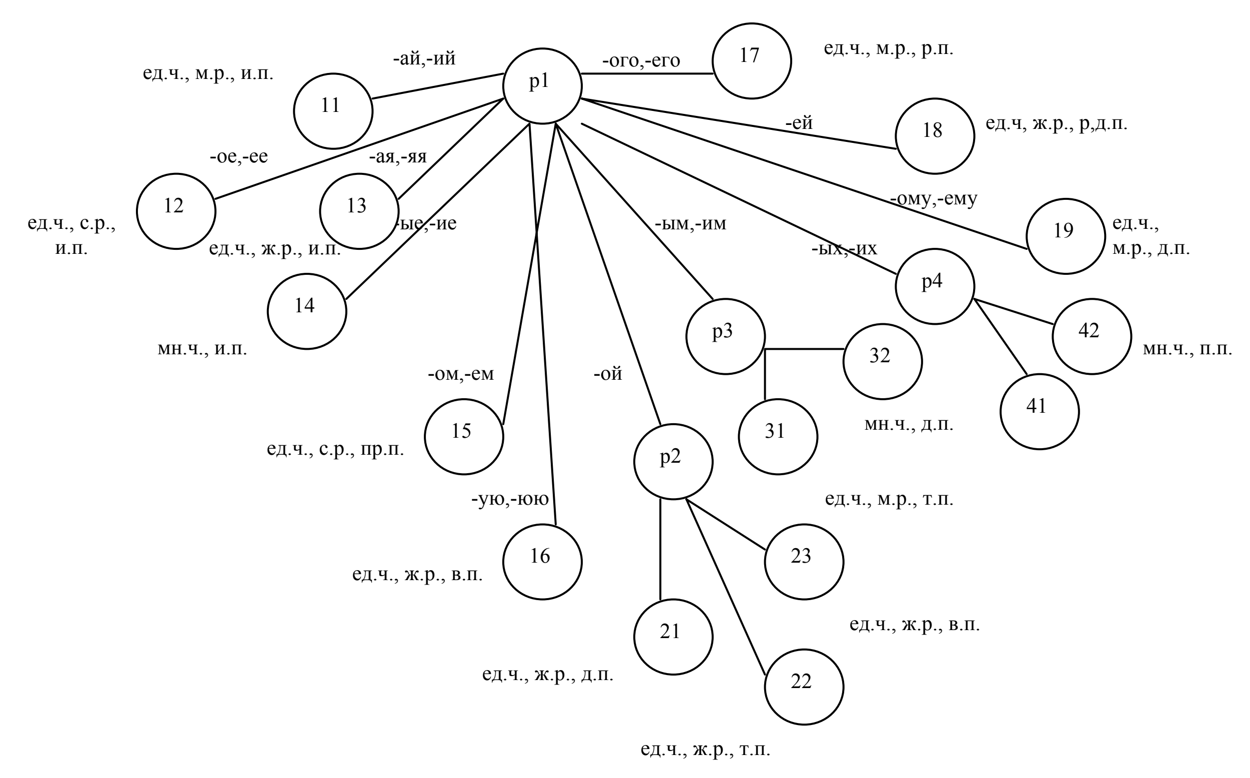


Рисунок 2 – Граф анализа словоформ адъективного склонения

Заключение

Таким образом, в настоящей работе предложен алгоритм получения всех характеристик словоформы с помощью графов составляющих ее морфем. Использование графов для получения грамматических характеристик словоформы демонстрирует общность алгоритмических методов на морфологическом и синтаксическом уровнях лингвистического анализа.

В дальнейшем планируется построение графов анализа словоформ глаголов и наречий. Полученные результаты могут применяться при автоматизированном переводе и в модуле морфологического анализа полнотекстовых информационно-поисковых систем с естественноязыковым интерфейсом.

Литература

1. Волкова И.А., Руденко Т.В. Формальные грамматики и языки. Элементы теории трансляции. – М.: Изд-во МГУ, 1999.
2. Зализняк А.А. Грамматический словарь русского языка. – М.: Русские словари, 2003.
3. SRILM – The SRI Language Modeling Toolkit. – Режим доступа: http://www.speech.sri.com/projects/srilm.
4. Manning C., Schutze H. Foundations of Statistical Language Processing. – The MIT Press, 1999.
5. Егошина А.А. Об одном способе построения статического словаря морфологического процессора // Материалы Седьмой Международной научно-технической конференции «Искусственный интеллект. Интеллектуальные и многопроцессорные системы – 2006». – Т. 2. – Таганрог: Изд-во ТРТУ. – 2006. – 404 с.
6. Панова М.В. Словообразование современного русского литературного языка. – М.: Наука, 1968.

С.М. Вороной, Г.А. Єгошина

Визначення граматичних характеристик словоформи методом графів
Для систем автоматизованого перекладу та морфологічного процесора повнотекстових інформаційно-пошукових систем з природномовним інтерфейсом запропонований алгоритм визначення всіх характеристик словоформ субстантивної та ад'єктивної відміни за допомогою графів морфем.

Статья поступила в редакцию 14.01.2008.